ISSN 1870-4069

Identificación de variables causales de la congestión de tráfico mediante la prueba de Granger

Ernesto De la Cruz-Nicolás, Hugo Estrada-Esquivel,
Alicia Martínez-Rebollar, Odette Alejandra Pliego-Martínez

Tecnológico Nacional de México
Centro Nacional de Investigación y Desarrollo Tecnológico,
México

{d21ce090, hugo.ee, alicia.mr, d21ce092}@cenidet.tecnm.mx

Resumen. La identificación de variables causales de la congestión de tráfico que surgen en las grandes ciudades mediante la prueba de Granger es fundamental para comprender las variables que contribuyen al origen de los puntos de congestión de tráfico vehicular en áreas urbanas. La prueba de Granger evalúa la causalidad entre dos series temporales, lo que permite identificar si una variable precede a otra en el tiempo y, por lo tanto, tiene un efecto causal sobre ella en el contexto de identificar que variables impactan en la congestión del tráfico. Al aplicar la prueba de Granger a datos de tráfico e incidencias viales, es posible determinar qué variables impactan de manera directa e indirecta a la congestión de tráfico, como la velocidad de los vehículos, las condiciones climáticas, la infraestructura vial o eventos específicos, pueden ser considerados como causas significativas de la congestión. Identificar las variables causales de la congestión de tráfico ofrece una serie de beneficios significativos como el desarrollo de estrategias efectivas de gestión del tráfico, promover la sostenibilidad ambiental, seguridad vial. La investigación en cuestión se suma a otras técnicas de la literatura que se dedican a identificar, correlacionar y relacionar variables causales de la congestión de tráfico, con el propósito de perfeccionar la precisión de los modelos predictivos en esta área.

Palabras clave: Causalidad, congestión, tráfico, impacto.

Identification of Causal Variables of Traffic Congestion Using the Granger Test

Abstract. The identification of causal variables of traffic congestion that arise in large cities through Granger causality testing is fundamental to understanding the factors contributing to the origin of traffic congestion points in urban areas. The Granger test assesses causality between two time series, allowing for the identification of whether one variable precedes another in time and thus has a causal effect on it in the context of identifying which variables impact traffic congestion. By applying the Granger test to traffic data and road incidents, it is possible to determine which variables directly and indirectly impact traffic congestion, such as vehicle speed, weather conditions, road infrastructure, or

specific events, which can be considered significant causes of congestion. Identifying the causal variables of traffic congestion offers a range of significant benefits such as the development of effective traffic management strategies, promoting environmental sustainability, and road safety. The research in question adds to other techniques in the literature dedicated to identifying, correlating, and relating causal variables of traffic congestion, with the purpose of refining the accuracy of predictive models in this area.

Keywords: Causality, congestion, traffic, impact.

1. Introducción

La congestión de tráfico es un fenómeno que aparece en las ciudades urbanizadas de todo el mundo, con consecuencias significativas en términos de tiempo de viaje perdido en los desplazamientos, afectación a la salud por el aumento de emisiones de gases de efecto invernadero, contaminación entre otros. En el intento por mitigar el problema de la congestión de tráfico, es fundamental comprender las variables subyacentes que contribuyen al origen de la congestión vehicular.

Existen en la literatura un extenso trabajo relacionado con modelos predictivos que han utilizado variables significativas para predecir congestiones de tráfico. La regresión lineal o múltiple, se han utilizado para predecir la congestión de tráfico en función de variables significativas como el flujo de tráfico, la velocidad promedio, las condiciones meteorológicas, la hora del día, eventos especiales y la densidad poblacional, entre otros. Estos modelos pueden proporcionar estimaciones cuantitativas de la probabilidad o el grado de congestión en diferentes momentos y ubicaciones [1, 2, 3].

Las redes neuronales artificiales utilizan las variables significativas que causan la congestión de tráfico como entradas para predecir la congestión de tráfico en función de múltiples factores, incluyendo el historial de tráfico, datos de sensores, eventos especiales y condiciones ambientales [4, 5, 6].

Los modelos de series temporales, se utilizan para predecir la congestión de tráfico en función de patrones temporales pasados. Estos modelos incorporan variables significativas como la hora del día, el día de la semana, días festivos y eventos especiales para prever futuros niveles de congestión con base en tendencias históricas [7, 8, 9].

Los modelos de simulación de tráfico, como VISSIM o SUMO, simulan la congestión de tráfico en función de diferentes variables, como la geometría de la carretera, la demanda de tráfico, los semáforos y las condiciones meteorológicas. Estos modelos predicen las congestiones de tráfico en tiempo real [10, 11, 12].

En este contexto, la importancia de identificar las variables causales de la congestión de tráfico es fundamental para incrementar la precisión y otras métricas en los modelos predictivos y otras técnicas para la caracterización, pronóstico y predicción de la congestión de tráfico. Las técnicas estadísticas ofrecen un enfoque poderoso para analizar la relación causal entre diferentes variables y la congestión de tráfico e identificar las variables de mayor significancia en el origen de la congestión de tráfico. La prueba de Granger, es una herramienta que permite evaluar si una serie temporal de una variable puede predecir de manera significativa otra serie temporal,

proporcionando así información valiosa sobre las variables causales detrás de la congestión de tráfico.

En este artículo, exploraremos la aplicación de la prueba de Granger para identificar las variables causales de la congestión de tráfico, a partir de una propuesta de un conjunto de variables relacionados con el tráfico e incidencias viales que impactan en el origen de la congestión de tráfico. A través de esta investigación, se contribuye a una comprensión más profunda de los factores que influyen en la congestión de tráfico, lo que a su vez puede proporcionar variables significativas de congestión de tráfico a los modelos predictivos para mejorar su precisión.

2. Trabajos relacionados

La rápida expansión de la urbanización en las grandes ciudades del mundo ha generado un aumento sin precedentes en el origen de múltiples puntos de congestión de tráfico, un problema constante con repercusiones significativas en la salud pública y el medio ambiente, como se destacan los hallazgos en el trabajo de [14]. La congestión de tráfico, al ser un desafío complejo y persistente, ha motivado una amplia gama de investigaciones orientadas al desarrollo de estrategias para mitigar este problema, tal como lo señala Yue [24].

Sin embargo, es fundamental comprender qué variables ejercen una mayor influencia en el origen de la congestión de tráfico, ya que esto no solo mejora la precisión de los modelos predictivos, sino que también optimiza otras técnicas y estrategias destinadas a mitigar la congestión de tráfico y se evita caer en la maldición de la multidimensionalidad que refiere al número de variables de un conjunto de datos, lo que puede dificultar el análisis y el modelado de manera efectiva. Esta maldición puede provocar problemas como el aumento de la complejidad computacional, la dispersión de los datos y el sobreajuste de los modelos de acuerdo al trabajo de [25].

La literatura existente ha llevado a cabo una serie de estudios para evaluar las variables que causan la congestión de tráfico. Por ejemplo, en el trabajo de Kardani-Yazd [13], encuentran relaciones instantáneas entre variables climáticas/no climáticas locales y el flujo de tráfico como variables causales. Por otro lado, en la investigación de Iro [16], identifican que un 66% de la congestión del tráfico se debe a factores humanos, mientras que un 34% se atribuye a condiciones físicas de las calles.

Además, se ha observado que aspectos como los accidentes automovilísticos y la duración de la atención a los mismos, como lo señala Chen [17] son variables que impactan. Mientras el área de calle per cápita, la propiedad de vehículos y las millas recorridas por vehículo, como lo describe Bian [18], juegan un papel importante en la congestión del tráfico. Mahona [19] también identifica a las cruces existentes en las calles como influencia a la congestión de tráfico. Asimismo, Yu [20] ha resaltado la importancia de las intersecciones de calles como una de las principales variables que ocasionan la congestión de tráfico.

El análisis de puntos de atracción como centros comerciales, escuelas, oficinas y tiendas, como se describe en el trabajo de Gullotta [21], también se ha revelado como relevante para comprender el origen de la congestión de tráfico. Rahman [22], utilizando el Modelado de Ecuaciones Estructurales, identificó el agrupamiento de ingresos y empleo como variables predominantes en la congestión del tráfico. Por otro

lado, Pi [23] sugiere que la combinación del diseño de las calles, eventos en las vías, condiciones de la calle y malos hábitos de los conductores, puede contribuir a la congestión del tráfico.

La identificación de variables causales de la congestión de tráfico ha permitido la creación de indicadores que evalúan la calidad de la infraestructura vial y la movilidad vehicular. En el trabajo de Jia [15], se analizan 15 indicadores relevantes y se examinan las variables como la densidad de población urbana, el Producto Interno Bruto, factores sociales, económicos y la oferta de transporte público, entre otros.

La congestión de tráfico es un fenómeno complejo influenciado por múltiples variables. La prueba de Granger permite examinar la relación causal entre variables y comprender cómo influyen en la congestión de tráfico.

La comprensión de las variables causales es crucial para construir modelos predictivos precisos de congestión de tráfico. La prueba de Granger proporciona información sobre qué variables pueden utilizarse como predictores para pronosticar la congestión de tráfico futura. La prueba de Granger permite evaluar la relación causal entre variables y validar hipótesis sobre las variables que contribuyen a la congestión de tráfico. Esto ayuda a confirmar o refutar suposiciones previas y a generar nuevas ideas para investigaciones futuras.

3. Metodología para identificar las variables causales de la congestión de tráfico mediante la prueba de Granger

La metodología para identificar las variables que causan la congestión de tráfico mediante la técnica estadística de Granger consta de cinco fases. La primera fase implica la definición del caso de estudio, donde se selecciona la zona geográfica que será objeto de análisis de la congestión de tráfico. En la segunda fase, se lleva a cabo la recolección de datos de tráfico e incidencias viales, donde se recopila información detallada y precisa sobre diversos aspectos relacionados con el flujo vehicular y los incidentes que afectan la circulación. En la tercera fase, se realiza el preprocesamiento de los datos de tráfico e incidencias viales. Durante esta etapa, los datos se transforman para mejorar su normalidad y se estandarizan debido a la diversidad de variables presentes.

La cuarta fase implica la implementación de la prueba de Granger a las variables de tráfico e incidencias viales con respecto a la congestión de tráfico (jamFactor). Aquí, se utiliza el método estadístico de Granger para identificar posibles relaciones causales entre estas variables. Finalmente, en la quinta fase, se lleva a cabo la validación de las variables causales de la congestión de tráfico encontradas mediante la prueba de Granger. Durante esta fase, se emplean las variables causales de la congestión de tráfico como entrada para el algoritmo predictivo de Random Forest, analizando su precisión en función de las variables de entrada. La metodología propuesta consiste en un conjunto de fases diseñadas para identificar las variables causales más significativas entre las 25 variables propuestas en esta investigación. Este enfoque complementa otras técnicas de identificación, relación y correlación de variables de mayor impacto relacionadas con la congestión de tráfico. A continuación, se detalla cada fase de la metodología.

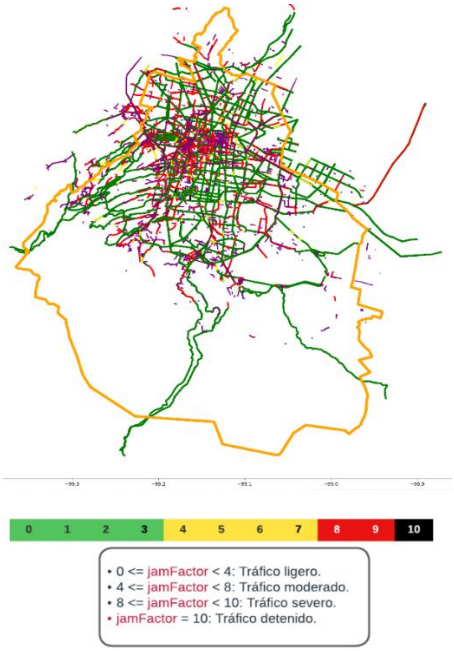

Fig. 1. Niveles de congestión de tráfico (jamFactor) de las calles de la Ciudad de México.

3.1. Caso de estudio

Una muestra de 6708 calles de las 16 alcaldías de la Ciudad de México será el caso de estudio para recopilar datos de tráfico e incidencias. En la Figura 1 se muestra la movilidad vehicular en las calles de la Ciudad de México a las 6:25 p.m. de un día martes. Las calles se representan con diferentes colores: verde, amarillo, rojo y negro, indicando distintos niveles de congestión de tráfico. Se destacan algunas calles en color morado, que representan aquellas con incidencias viales. El jamFactor oscila entre 0 y 10, donde un valor cercano a 0 indica una congestión de tráfico mínima, lo que implica un flujo vehicular ligero con una velocidad de desplazamiento adecuada. Por el contrario, un valor de jamFactor cercano a 10 indica una congestión de tráfico severa, con una velocidad de desplazamiento considerablemente reducida, mostrados en la Figura 1.

3.2. Recolección de datos de tráfico e incidencias viales

La recolección de datos de tráfico e incidencias viales se llevó a cabo durante un período que abarcó desde el 1 de septiembre hasta el 31 de diciembre de 2023, lo que representó un total de 122 días de extracción de datos. La recolección de datos se hizo mediante la aplicación de Here Maps, la cual fue configurada utilizando un script en Python para recopilar información en intervalos de solicitud de cada 5 minutos a lo largo de las 24 horas del día. En total, se logró reunir un conjunto de datos denominado

Tabla 1. Variables de tráfico e incidencias viales (Tráfico-Incidencias).

Conjunto de datos	Variable	Descripción
Tráfico	jamFactor	Describe los niveles de tráfico en un escala del 0 al 10.
	Bbox_Traffic	Cuadro delimitador donde se origina el tráfico.
	Free_Flow	Velocidad de desplazamiento sin ninguna obstrucción.
	Length	Longitud de la calle.
	Number_Segments	Número de segmentos que contiene la calle.
	Speed	La velocidad esperada.
	Speed_Uncapped	Velocidad sin restricción.
	Daily_Traffic	Día en que se monitorea el tráfico.
	Day_Off_Traffic	Indica si el día es festivo en cuanto a tráfico.
	Traffic_Hour	Hora del monitoreo de tráfico.
	Traffic_Minute	Minuto del monitoreo de tráfico.
	Traffic_Day_Number	Día de monitoreo de tráfico (1= lunes, 2=martes…7=domingo).
	Work_Day_Traffic	Monitoreo del tráfico en semana laboral o fin de semana
Incidencias	Bbox_Incident	Cuadro delimitador donde se origina la incidencia.
	Road_Closed	Calle cerrada por incidencia (0 abierto y 1 cerrado).
	Criticality	Gravedad de la incidencia.
	Type	Describe el tipo de incidencia.
	Incident_Day	Día del monitoreo de la incidencia.
	Day_Off_Incident	Indica si el día es festivo en cuanto a incidencia.
	Incident_Hour	Hora del monitoreo de la incidencia.
	Incident_Month	Mes del monitoreo de la incidencia.
	Incident_Minute	Minuto del monitoreo de la incidencia.
	Incident_Day_Number	Día de incidencia (1= lunes, 2=martes…7=domingo).
	Work_Day_Incident	Incidencia por laboral o fin de semana (1 o 0 respectivamente).
	Category_Number	Categoría de la incidencia.
	Time	Tiempo en minutos.

Tráfico-Incidencias con una impresionante cantidad de 5, 000, 000 registros de datos, los cuales incluyen 26 variables.

Las 26 variables relacionados con el tráfico e incidencias viales recopiladas se describen en la tabla 1.

3.3. Preprocesamiento de datos de tráfico e incidencias viales

El preprocesamiento de datos de tráfico e incidencias viales se llevó a cabo mediante dos pasos para mejorar la calidad y uniformidad de los datos. Esto incluyó tanto la transformación de datos para mejorar la normalidad como la estandarización para abordar las diferentes escalas de cada variable. A continuación, se describen en detalle los pasos realizados:

Paso 1, transformar los datos: La transformación se hizo mediante la raíz cuadrada, el z-score y la transformación cúbica, aplicadas mediante las ecuaciones 1, 2 y 3, respectivamente, en el conjunto de datos de Tráfico-Incidencias:

$$Variable\ normalizada = \sqrt{Variable\ original}, \tag{1}$$

$$z = \frac{(x-\mu)}{\sigma}, \tag{2}$$

dónde: x Es el valor individual de la variable, μ Es la media de la distribución de la variable y σ Es la desviación estándar de la distribución de la variable:

$$Variable\ normalizada = \sqrt[3]{Variable\ original}. \quad (3)$$

Los hallazgos derivados de la normalización del conjunto de datos de Tráfico-Incidencias señalan que la aplicación de la transformación cúbica resulta en una notable mejora en la normalidad de este conjunto de datos.

Paso 2, estandarizar los datos: Se realizó la estandarización el conjunto de datos de Tráfico-Incidencias, en una escala de 0 a 1. Esta estandarización se efectuó utilizando la ecuación número 4:

$$x\_norm = \frac{(x - x\_min)}{(x\_max - x\_min)}, \quad (4)$$

dónde:

x Es el valor original que se desea normalizar, x_min Es el valor mínimo en el conjunto de datos original y x_max Es el valor máximo en el conjunto de datos original.

La estandarización produjo que todas las variables del conjunto de datos de Tráfico-Incidencias fueran ajustadas a un intervalo de valores que oscila entre 0 y 1. Este proceso garantiza que todas las variables estuvieran en una escala uniforme y comparable entre sí.

3.4. Implementación de la prueba de Granger a las variables de tráfico e incidencias viales

La prueba de causalidad de Granger es una técnica econométrica utilizada para determinar si existe una relación de causalidad entre dos variables de series de tiempo. La idea principal detrás de esta prueba es evaluar si los valores pasados de una variable X pueden ayudar a predecir los valores futuros de otra variable Y, más allá de la información contenida en los valores pasados de Y.

La implementación de la prueba de Granger a las variables del conjunto de datos Tráfico-Incidencias implica varios pasos clave. Aquí se describen estos pasos de manera detallada:

Paso 1 seleccionar variables: se utilizan las 25 variables de tráfico e incidencias viales descritas en la tabla 1, exceptuando la variable jamFactor.

Paso 2 establecer las hipótesis: para identificar la causalidad entre dos variables mediante la prueba de Granger se establecen las siguientes hipótesis:

> ***Hipótesis nula (H0):*** No hay causalidad entre las dos variables. En este caso la serie temporal de una variable no puede predecir la otra variable.
>
> ***Hipótesis alternativa (H1):*** Existe causalidad entre las dos variables. La serie temporal de una variable puede predecir la otra variable.

Si el valor p es menor que el nivel de significancia, en este caso de 0.05, entonces se rechaza la hipótesis nula y concluimos que hay evidencia suficiente para sugerir que la serie temporal de una variable influye en la otra variable.

Si el valor p es mayor que el nivel de significancia, no se rechaza la hipótesis nula y no hay suficiente evidencia para afirmar que una variable causa la otra.

Paso 3 obtener el valor p entre las variables con la prueba de causalidad de Granger: para calcular el valor p entre las variables descritas en la tabla 1 y el factor de atasco (jamFactor), donde las variables de la tabla 1 serán referidas como X y el jamFactor como Y, se emplea la prueba de causalidad de Granger. Inicialmente, se ajusta un modelo de regresión lineal que incorpora los retrasos del par de variables. Luego, se lleva a cabo una prueba de hipótesis para evaluar la significancia de los coeficientes relacionados con los retrasos de una variable en la predicción de la otra. Se aplica la ecuación 5 para calcular el p-valor en la prueba de Granger:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \gamma 1 X_{t-1} + \varepsilon t \,, \tag{5}$$

dónde:

Y_t Es el registro actual de la serie temporal Y, X_t Es el registro actual de la serie temporal X, Y_{t-1} Es el registro pasado de la serie temporal Y (retraso 1), X_{t-1} Es el registro pasado de la serie temporal X (retraso 1), α Es el intercepto, β_1 Es el coeficiente asociado al retraso de la serie temporal Y, $\gamma 1$ Es el coeficiente asociado al retraso de la serie temporal X y εt Es el término de error.

Se utilizó la clase grangercausalitytests del paquete estadístico statsmodels en Python para llevar a cabo la prueba de causalidad de Granger y determinar el valor de p entre las variables mencionadas en la tabla 1 con la variable jamFactor, la cual representa la congestión de tráfico. La prueba de Granger se implementó con valor de retraso de 3. Se estableció un nivel de significancia de 0.05 para identificar las variables con un valor de p significativamente menor a 0.05. El conjunto de datos "Trafico-Incidencias" se dividió en dos subconjuntos (TI1 para los meses de septiembre y octubre, y TI2 para noviembre y diciembre), con el fin de realizar dos pruebas separadas y determinar si los resultados muestran consistentemente las mismas variables de mayor impacto. Aquellas variables que cumplen con este criterio se detallan en la tabla 2 como las variables que poseen una influencia estadísticamente significativa en la congestión de tráfico representada por jamFactor.

Las variables presentadas en la tabla 2 muestran valores p inferiores al umbral de significancia de 0.05, en las dos pruebas realizadas, se observan fluctuaciones en los valores de "p" a pesar de que las variables analizadas se mantienen constantes, lo que indica una relación causal directa con la variable jamFactor, de acuerdo a la prueba de causalidad de Granger. Sin embargo, las variables ausentes en la tabla 2 están vinculadas con otras variables que no tienen una conexión directa con jamFactor, pero están interrelacionadas con otras variables, como se muestra en la figura 2. Estas variables ejercen un impacto indirecto en la congestión del tráfico (jamFactor). Por otro lado, las variables no representadas en la figura 2, se consideran independientes dado que no presentan ningún tipo de relación con otras variables.

Las variables que inciden en jamFactor muestran una relación causal con la congestión de tráfico, como se detalla en la tabla 2. La similitud en las series de tiempo entre Number_Segments y jamFactor, así como entre Speed y jamFactor se muestran en la figura 3 y 4 respectivamente.

La aplicación de la prueba de Granger ha revelado que las variables ***Type, Speed_Uncapped, Speed, Number_Segments y Free_Flow*** son las más significativas en la causalidad de la congestión de tráfico, dentro del conjunto de 25 variables consideradas y detalladas en la tabla 1. Este resultado indica que estas variables tienen

Tabla 2. Variables con valores de p inferiores al nivel de significancia de 0.05.

X	Y	Valor F de TI1	Valor P de TI1	Valor F de TI2	Valor P de TI1
Type	jamFactor	4.5	0.03	3.9	0.045
Speed_Uncapped	jamFactor	9.6	0.01	4.7	0.029
Speed	jamFactor	8.9	0.002	10.5	0.001
Number_Segments	jamFactor	6.1	0.01	18.1	0.002
Free_Flow	jamFactor	6.1	0.01	7.6	0.005

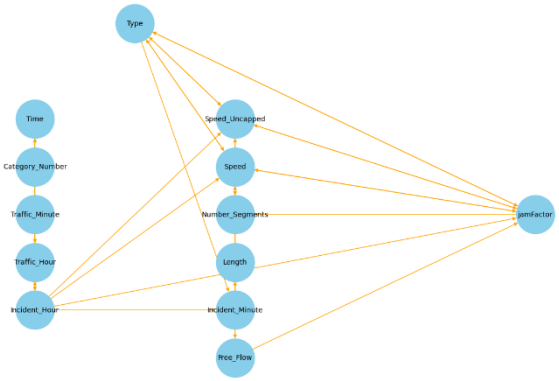

Fig. 2. Causalidad entre todas las variables, donde cada variable muestra una relación de causalidad con una o más variables.

un impacto estadísticamente significativo en la predicción de la congestión vehicular, lo que sugiere su relevancia en el análisis y la gestión del tráfico.

3.5. Validación de las variables causales de congestión de tráfico identificadas mediante la prueba de Granger

La validación de las variables causales de congestión de tráfico identificadas mediante la prueba de Granger en la sección 3.4 consiste en la implementación del algoritmo Random Forest para evaluar la importancia de las variables identificadas en la predicción de la congestión de tráfico y para validar su causalidad. Se ha seleccionado el algoritmo de Random Forest debido a que es robusto frente al sobreajuste, lo que lo hace adecuado para evitar problemas comunes al trabajar con conjuntos de datos grandes, como es el caso de la congestión vehicular. Además, Random Forest proporciona una medida de importancia para cada variable, lo que facilita la identificación de aquellas que tienen un mayor impacto en la predicción de la congestión del tráfico. Su capacidad para manejar tanto variables categóricas como numéricas sin requerir transformaciones adicionales lo hace particularmente útil en este escenario heterogéneo. A continuación se detallan los pasos:

Paso 1, establecer el conjunto de datos: se eligen aleatoriamente 8,983 registros de los datos recopilados de tráfico e incidentes viales correspondientes al mes de enero de 2024, previamente equilibrados mediante la técnica de SMOTE (Técnica de Sobremuestreo Sintético para la Minimización de la Clase Minoritaria). Estos registros contienen información sobre cinco variables específicas: Type, Speed_Uncapped, Speed, Number_Segments y Free_Flow.

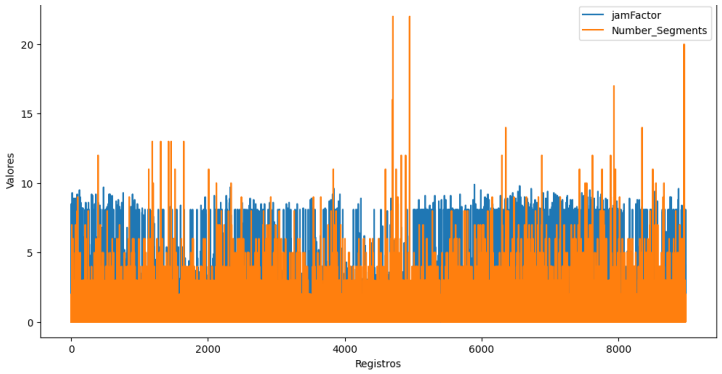

Fig. 3. Causalidad entre las variable Number_Segments con respecto a jamFactor.

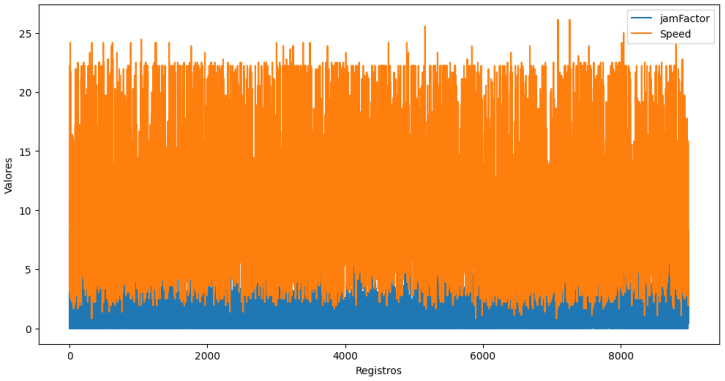

Fig. 4. Causalidad entre las variable speed con respecto a jamFactor.

Paso 2, establecer la variable objetivo: la variable objetivo en este trabajo de investigación es la congestión de tráfico representada por la variable jamFactor.

Paso 3, dividir los datos en conjuntos de entrenamiento y prueba: de acuerdo a la literatura para el entrenamiento de modelos predictivos se determina el 80% (7186 registros) de los datos para el entrenamiento y el 20% (1797 registros) para la prueba.

Paso 4, entrenar el modelo: para entrenar el modelo, importamos la clase RandomForestClassifier de la biblioteca sklearn en Python, específicamente para realizar una tarea de clasificación. Establecemos los valores de los hiperparámetros mediante el uso del algoritmo de Random Search, y estos valores quedaron definidos de la siguiente manera: n_estimators se fija en 100 para definir el número de árboles en el bosque; max_depth se establece en None para permitir que los árboles crezcan sin restricciones de profundidad máxima; min_samples_split se define como 10, estableciendo el número mínimo de muestras requeridas para dividir un nodo interno; min_samples_leaf se fija en 5 para determinar el número mínimo de muestras permitidas en una hoja; y bootstrap se establece en verdadero para permitir el muestreo con reemplazo durante la construcción de los árboles.

Una vez entrenado el modelo se muestra la matriz de confusión en la figura 5 que permite visualizar el desempeño del algoritmo de clasificación entrenado comparando las predicciones del modelo con los valores verdaderos.

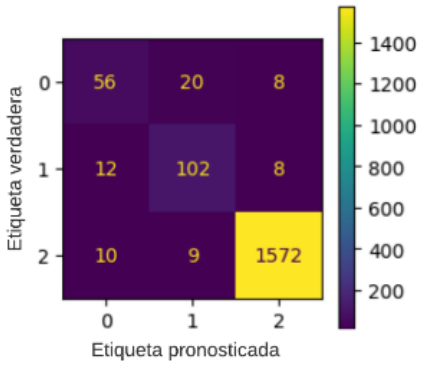

Fig. 5. Matriz de confusión del algoritmo Random Forest con las variables más relevantes.

La matriz de confusión de la figura 5 proporciona una descripción detallada del desempeño del modelo de clasificación de congestión de tráfico, en este caso, para tres clases: "Amarillo", "Rojo" y "Verde". A continuación se explican los resultados mostrados en la matriz de confusión:

Verdaderos Positivos (TP): para la clase "Amarillo", hay 56 casos en los que el modelo predijo correctamente "Amarillo". Para la clase "Rojo", hay 102 casos en los que el modelo predijo correctamente "Rojo". Para la clase "Verde", hay 1572 casos en los que el modelo predijo correctamente "Verde".

En el caso de los Falsos Positivos (FP): Para la clase "Amarillo", hay 20 casos en los que el modelo predijo incorrectamente "Amarillo" cuando en realidad era "Rojo" (12 casos) o "Verde" (8 casos). Para la clase "Rojo", hay 12 casos en los que el modelo predijo incorrectamente "Rojo" cuando en realidad era "Amarillo" (8 casos) o "Verde" (4 casos). Para la clase "Verde", hay 9 casos en los que el modelo predijo incorrectamente "Verde" cuando en realidad era "Amarillo" (8 casos) o "Rojo" (1 caso).

La exactitud global del modelo en el conjunto de prueba es del 96.27%. Esto indica la proporción de predicciones correctas en general.

4. Resultados

Los resultados de la prueba de Granger indicaron que cinco variables específicas, a saber, Type, Speed_Uncapped, Speed, Number_Segments y Free_Flow, mostraron una causalidad significativa con respecto al jamFactor, que se utiliza como variable para medir la congestión de tráfico.

Esto implica que el tipo de incidente (Type) que ocurra en las calles incide directamente en el surgimiento de congestiones de tráfico. Por lo tanto, en las vías, cuando se registra algún percance, la congestión resultante es dinámica, dependiendo del tipo de incidencia.

La velocidad esperada (Speed) en las calles es alta, es decir, cuando el flujo de vehículos es fluido y pueden desplazarse a una velocidad adecuada, la congestión tiende a ser mínima. Sin embargo, cuando esta velocidad disminuye debido a factores como

el aumento del volumen de tráfico y otras variables relacionadas la velocidad esperada es baja.

La velocidad sin restricciones (Speed_Uncapped) describe la velocidad constante que se pueden visualizar de las calles y que ayudan a minimizar los puntos de congestión y mantener la eficiencia de la movilidad vehicular.

El número de segmentos en la calle (Number_Segments) se vuelve significativo, cuando la calle tiene muchas intersecciones y aumenta las probabilidades de que los vehículos deban detenerse o reducir su velocidad para ceder el paso a otros vehículos que ingresan desde diferentes direcciones, lo que contribuye a la congestión.

Además, el flujo libre (Free_Flow) de la calle, que representa la capacidad de los vehículos para moverse sin obstáculos ni restricciones, también juega un papel crucial en la congestión de tráfico. Estas relaciones descritas anteriormente son estadísticamente significativas, según los valores de p obtenidos en el análisis y que impactan en la congestión de tráfico.

Por otro lado, las otras variables del conjunto de datos no mostraron una causalidad significativa con el jamFactor. Sin embargo, es importante destacar que estas variables se encuentran relacionadas con otras variables que afectan indirectamente la congestión de tráfico, aunque no sean directamente causales con el jamFactor.

5. Conclusiones y discusión

La validación de las 5 variables más significativas mediante un algoritmo de clasificación Random Forest demuestra la capacidad predictiva de las cinco variables identificadas. El alto nivel de precisión obtenido en la clasificación de diferentes niveles de congestión respalda la relevancia de estas variables en la predicción de la congestión de tráfico.

Identificar las variables clave que influyen en la congestión de tráfico puede permitir a las autoridades de transporte diseñar estrategias más efectivas para mitigar la congestión y mejorar la movilidad en las ciudades.

La selección de variables específicas mediante el uso de la aplicación de mapas (Here Maps) propuestas en este trabajo de investigación y el enfoque en un método estadístico particular pueden ser reforzadas con otras técnicas de análisis de datos. Por ejemplo, el estudio podría beneficiarse de la aplicación de técnicas como el Análisis de Componentes Principales, el Análisis Factorial, la Correlación de Spearman, Modelos de Ecuaciones Estructurales, entre otros.

Estas técnicas podrían ayudar a explorar más a fondo las relaciones entre las variables y la congestión de tráfico, identificando posibles interacciones complejas o relaciones no lineales que podrían haber sido pasadas por alto en el análisis inicial. Además, la incorporación de variables adicionales que podrían influir en la congestión de tráfico, como datos de cultura de manejo, dimensiones de la infraestructura vial, nivel de estrés de los usuarios, variables socioeconómicas entre otros, podrían proporcionar una comprensión más completa de las variables que contribuyen a la congestión de tráfico.

Referencias

1. Liu, Y., Liu, C., Zheng, Z.: Traffic congestion and duration prediction model based on regression analysis and survival analysis. Open Journal of Business and Management, vol. 08, no. 02, pp. 943–959 (2020). DOI: 10.4236/ojbm.2020.82059.
2. Tamir, T.S., Xiong, G., Li, Z., Tao, H., Shen, Z., Hu, B., Menkir, H.M.: Traffic congestion prediction using decision tree, logistic regression and neural networks. IFAC-PapersOnLine, vol. 53, no. 5, pp. 512–517 (2020). DOI: 10.1016/j.ifacol.2021.04.138.
3. Fahs, W., Chbib, F., Rammal, A., Khatoun, R., Attar, A.E., Zaytoun, I., Hachem, J.: Traffic congestion prediction based on multivariate modelling and neural networks regressions. Procedia Computer Science, vol. 220, pp. 202–209 (2023). DOI: 10.1016/ j.procs.2023.03.028.
4. Nuli, S., Vikranth, N., Gupta, K.A.: Real-time traffic prediction using neural networks. In: IOP Conference Series. Earth and Environmental Science, vol. 1086, no. 1, pp. 012029 (2022). DOI: 10.1088/1755-1315/1086/1/012029.
5. Nuli, S., Vikranth, N., Gupta, K.A.: Real-time traffic prediction using neural networks. In: IOP Conference Series. Earth and Environmental Science, vol. 1086, no. 1, pp. 012029 (2022). DOI: 10.1088/1755-1315/1086/1/012029.
6. Iyer, P.R., Iyer, S.R., Ramesh, R., Subramanya, K.N.: Adaptive real time traffic prediction using deep neural networks. In: IAES International Journal of Artificial Intelligence (IJ-AI), vol. 8, no. 2, pp.107 (2019). DOI: 10.11591/ijai.v8.i2.pp107-119.
7. Rajalakshmi-Vaidyanathan, G.: Hybrid time-series forecasting models for traffic flow prediction. PROMET - TrafficandTransportation, vol. 34 no. 4, pp. 537–549 (2022). DOI: 10.7307/ptt.v34i4.3998.
8. Dadashova, B., Li, X., Turner, S., Koeneman, P.: Multivariate time series analysis of traffic congestion measures in urban areas as they relate to socioeconomic indicators. Socio-Economic Planning Sciences, vol. 75, no. 100877 (2021). DOI: 10.1016/j.seps.2020. 100877.
9. Van der-Bijl, B., Gijsbertsen, B., van Loon, S., Reurich, Y., de Valk, T., Koch, T., Dugundji, E.: A comparison of approaches for the time series forecasting of motorway traffic flow rate at hourly and daily aggregation levels. Procedia Computer Science, vol. 201, pp. 213–222 (2022). DOI: 10.1016/j.procs.2022.03.030.
10. Wu, Y., Lin, Y., Hu, R., Wang, Z., Zhao, B., Yao, Z.: Modeling and simulation of traffic congestion for mixed traffic flow with connected automated vehicles: A cell transmission model approach. Journal of Advanced Transportation, vol. 2022, pp. 1–20 (2022). DOI: 10.1155/2022/8348726.
11. Hu, W., Wang, H., Qiu, Z., Yan, L., Nie, C., Du, B.: An urban traffic simulation model for traffic congestion predicting and avoiding. Neural Computing and Applications, vol. 30, no. 6, pp. 1769–1781 (2018). DOI: 10.1007/s00521-016-2785-7.
12. Dorokhin, S., Artemov, A., Likhachev, D., Novikov, A., Starkov, E.: Traffic simulation: an analytical review. In: IOP Conference Series. Materials Science and Engineering, vol. 91, no. 1, pp. 012058 (2020). DOI: 10.1088/1757-899x/918/1/012058.
13. Kardani-Yazd, N., Kardani-Yazd, N., Mansouri-Daneshvar, M.R.: A rapid method for evaluating the variables affecting traffic flow in a touristic road, Iran. Environmental Systems Research, vol. 8, no. 34 (2019). DOI: 10.1186/s40068-019-0162-0.
14. Afrin, T., Yodo, N.: A survey of road traffic congestion measures towards a sustainable and resilient transportation system. Sustainability, vol. 12, no. 11, pp. 4660 (2020). DOI: 10.3390/su12114660.

15. Jia, X.: Analysis on influencing factors of traffic congestion in Chongqing and study on countermeasures: Empirical analysis based on principal component analysis. Atlantis Press International BV, pp. 814–822 (2023). DOI: 10.2991/978-94-6463-200-2_84.
16. Iro, S., Pat-Mbano, E.C.: Causes of traffic congestion: A study of owerri municipal area of IMO state. American Journal of Environmental Sciences, vol. 18, no. 3, pp. 52–60 (2022). DOI: 10.3844/ajessp.2022.52.60.
17. Chen, L., Shi, J., Cheng, M., Zhu, H., Sun, L.: Characteristics of urban road non-recurrent traffic congestion based on floating car data. In: 4th International Conference on Electronic Information Technology and Computer Engineering, pp. 120–126 (2020). DOI: 10.1145/3443467.3443740.
18. Bian, C., Yuan, C., Kuang, W., Wu, D.: Evaluation, classification, and influential factors analysis of traffic congestion in Chinese cities using the online map data. Mathematical Problems in Engineering, pp. 1–10 (2016). DOI: 10.1155/2016/1693729.
19. Mahona, J., Mhilu, C., Kihedu, J., Bwire, H.: Factors contributing to traffic flow congestion in heterogenous traffic conditions. International Journal for Traffic and Transport Engineering, vol. 9, no. 2, pp. 238–254 (2019). DOI: 10.7708/ijtte.2019.9(2).09.
20. Yu, J., Wang, L., Gong, X.: Study on the status evaluation of urban road intersections traffic congestion base on AHP-TOPSIS modal. Procedia, Social and Behavioral Sciences, vol. 96, pp. 609–616 (2013). DOI: 10.1016/j.sbspro.2013.08.071.
21. Gullotta, G., Loret, E., Stewart, C., Sarti, F.: Traffic attractors and congestion in the urban context, the case of the city of Rome. Journal of Geographic Information System, vol. 12, no. 6, pp. 545–559 (2020). DOI: 10.4236/jgis.2020.126032.
22. Rahman, M.M., Najaf, P., Fields, M.G., Thill, J.-C.: Traffic congestion and its urban scale factors: Empirical evidence from American urban areas. International Journal of Sustainable Transportation, vol. 16, no. 5, pp. 406–421 (2022). DOI: 10.1080/15568318.2021.1885085.
23. Pi, M., Yeon, H., Son, H., Jang, Y.: Visual Cause Analytics for Traffic Congestion. In: IEEE Transactions on Visualization and Computer Graphics, vol. 27. No. 3, pp. 2186–2201 (2021). DOI: 10.1109/tvcg.2019.2940580.
24. Yue, W., Li, C., Chen, Y., Duan, P., Mao, G.: What is the root cause of congestion in urban traffic networks: Road infrastructure or signal control?. In: IEEE Transactions on Intelligent Transportation Systems: A Publication of the IEEE Intelligent Transportation Systems Council, vol. 23, no. 7, pp. 8662–8679 (2022). DOI: 10.1109/tits.2021.3085021.
25. Bellman, R., Kalaba, R.E.: Dynamic programming and modern control theory. New York: Academic Press, vol. 81 (1965)